SCIENCE
Sociedade para o Desenvolvimento da Pesquisa Científica

INTEGRATION OF HOUSEHOLD SURVEYS AND OTHER DATA SOURCES

DENISE BRITZ DO N. SILVA

SCIENCE

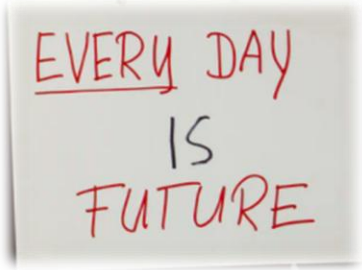
SOCIEDADE PARA O DESENVOLVIMENTO DA PESQUISA CIENTÍFICA

SEMINARIO REGIONAL SOBRE INNOVACIONES Y DESAFÍOS EN ENCUESTAS DE HOGARES - CEPAL

21 Nov 2024

THE CONTOURS - EVOLVING CHALLENGES

- Increasing demand for timely and disaggregated statistics
- Data revolution and the AI race
- Agenda 2030 and SDG Indicators



- Need to optimise resources for the statistical production process
- Data ecosystem for official statistics:
  NSOs, public sector, private sector + citizen-generated data
- Relevance of integrated statistical and geospatial information

DATA INTEGRATION

- Process to combine data from disparate sources into meaningful and valuable information (UN.ESCAP, 2020)
- Combination of data from different data sources to produce new data sets that are the basis for statistics or research

Data sources

- **Statistical surveys (probability samples)** and Census
- Administrative registers
- Nonprobability surveys
- Geospatial information
- Big data and citizen-generated data

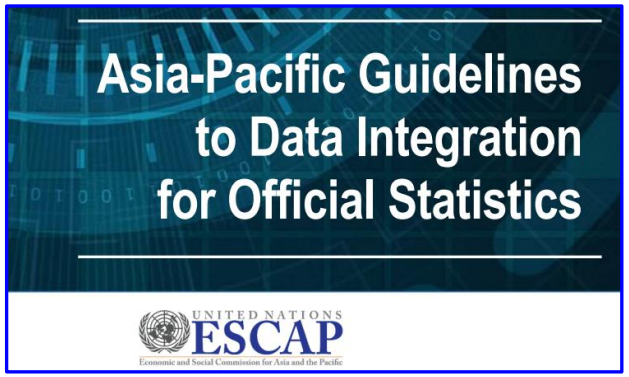
DATA INTEGRATION EFORTS ARE EVERYWHERE

SCIENCE

Seminario Integración de Fuentes de Datos y Estadísticas Oficiales - México, May 2024 - **INEGI and CEPAL**

International Seminar on Integrating Household Surveys with Diverse Data Sources

19-21 Nov 2024- UNSD – NBS CHINA



DATA INTEGRATION IN THE CONTEXT OF SURVEYS

Data Integration for, with, within and between household surveys

- Multiple data sources for the design and estimation phases
- Combining data and/or estimates from multiple surveys
- Use of a non-random data set to improve the statistical value of estimates from a random sample (Tam, 2020)
- Combining probability and nonprobability samples
- Data to supplement or replace the sample survey
- Multisource statistics (Waal et al., 2020)
- Statistical methods: record linkage, matching, modelling, spatial statistics, survey methods (finite population framework)

DATA INTEGRATION AND THE SURVEY PROCESS

- Datasets combined to create or maintain survey frames
- Several uses of auxiliary data:
  - Survey and sample design
  - Measures of variability for design variables and/or size measures
  - Edit and imputation (accuracy checks, edit rules or modelling)
  - Non-response and coverage adjustments
  - Weighting, calibration and poststratification
  - Model-based estimation procedures
  - Validation of survey estimation and quality assurance

POSITIONING HOUSEHOLD SURVEYS FOR THE NEXT DECADE

- Multi-frame designs: to capture groups underrepresented in census frames
- Spatial sampling: to establish the main sampling frame
- Responsive and adaptive sampling design: to deal with surveys' complexities and declining response rates
- Improve interoperability and integration of household surveys with other data sources

Statistical Journal of the IAOS 38 (2022) 923–946
DOI 10.3233/SJI-220042
Calogero Carletto, Haoyi Chen, Talip Kilica and Francesca Perucci

SURVEY METHODS FOR DATA INTEGRATION

SCIENCE

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Research Strand 7: Data integration

Statistical Science
2017, Vol. 32, No. 2, 293–312
DOI: 10.1214/16-STS584
© Institute of Mathematical Statistics, 2017

Combining Survey Data with Other Data Sources

Sharon L. Lohr and Trivellore E. Raghunathan

ESTUDIOS ESTADÍSTICOS 97

Desagregación de datos en encuestas de hogares

Metodologías de estimación en áreas pequeñas

Isabel Molina
CEPAL(2019)

isi International Statistical Review

International Statistical Review (2021), 89, 2, 382–401 doi:10.1111/insr.12434

Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference

Jae-Kwang Kim[1] and Siu-Ming Tam[2]

University of Essex · University of Southampton · Economic and Social Research Council · NCRM National Centre for Research Methods · Office for National Statistics · UCL · Warwick · Manchester · City University of London · National Centre for Social Research · LSE · Ulster University · Unil · Ipsos

Journal of Statistical Planning and Inference

An efficient approach for statistical matching of survey data through calibration, optimal transport and balanced sampling

Raphaël Jauslin *, Yves Tillé
University of Neuchâtel, Institute of statistics, Av. de Bellevaux 51, Neuchâtel, 2000, Switzerland

The Survey Statistician, 2022, Vol. 85, 19–29.

A gentle introduction to data integration in survey sampling

Jae Kwang Kim

Department of Statistics, Iowa State University, jkim@iastate.edu

DATA INTEGRATION WITH BRAZILIAN SURVEYS

- Weighting for a non-probability sample carried out in the pandemic scenario

  Marcelo Pitta (Brazilian Network Information Center - NIC.br), Pedro Silva

- Transformed Fay-Herriot Models for Quarterly Small Area Estimates of Extreme Poverty Rates in Brazil

  Guilherme Jacob, Nikos Tzavidis, Ángela Luna-Hernández and Pedro Silva (ENCE/IBGE and Southampton University)

- Time series models for repeated household sample surveys

  Caio Gonçalves (Fundação João Pinheiro), Luna Hidalgo (IBGE), Denise Silva and Jan van den Brakel (Statistics Netherlands)

WEIGHTING FOR A NON-PROBABILITY SAMPLE CARRIED OUT IN THE PANDEMIC SCENARIO

- NIC.br is responsible for planning, evaluating and monitoring the use ICT in Brazil
- Household Survey on the Use of ICT in Brazilian Households conducted annually - F2F interviewing
- During the pandemic, household survey was suspended
- ICT COVID-19 web panel survey
- Quota sample based on region, sex, age group, SES and education
- ~91k panel members contacted and ~2.5k responded
- Fieldwork period: June 23rd to July 8th, 2020

WEIGHTING FOR ICT COVID-19 WEB PANEL SURVEY

- Pseudo-weight estimation method to address selection biases of the ICT Panel COVID-19
- The 2019 ICT Household Survey was used as a reference sample for calculating pseudo-weights
- Inverse probability weighting - Elliott and Valliant (2017)
- Timetable from planning to publication: less than two months
- Other surveys using the web panel approach: three during the pandemic and three on new topics
- Survey methodology, challenges and accomplishments

https://cetic.br/en/publicacao/painel-tic-covid-19/

Elliott, Michael R., and Richard Valliant. 2017. Inference for Nonprobability Samples. Statistical Science 32 (2): 249–64. https://doi.org/10.1214/16-STS598

TRANSFORMED FAY-HERRIOT MODELS FOR QUARTERLY SMALL AREA ESTIMATES OF EXTREME POVERTY RATES IN BRAZIL

- IBGE produces annual estimates of extreme poverty rates at the national and state levels using its main household survey (PNADC)
- **SAE aim**: quarterly estimates (2013-2022) of extreme poverty rates at the Geographic Strata level (146 spatial domains)
- **SAE method**: Transformed Fay-Herriot models for proportions with area-time random effects
- **Auxiliary data**: administrative register of social program beneficiaries in Brazil, area indicator, indicator for quarters, AARR poverty rate and interactions

TRANSFORMED FAY-HERRIOT MODELS FOR QUARTERLY SMALL AREA ESTIMATES OF EXTREME POVERTY RATES IN BRAZIL

- Several transformations were tested
- SAE model represents spatiotemporal characteristics of Brazilian poverty from 2013 to 2022 (previously limited to the analysis of the state-level estimates produced annually by IBGE)
- On average, the MSEs of the model-based estimates were around 30% of the direct variances
- More disaggregated and more frequent statistical outputs

https://www.even3.com.br/anais/sinape2024/819122-beta-binomial-area-level-model-for-quarterly-estimates-of-extreme-poverty-rates-in-brazil/

TIME SERIES MODELS FOR REPEATED SURVEYS

- Repeated surveys collect data at several points in time
- **Repeated sample surveys: rotating panels**
  - Some units are retained on the sample from one occasion to the next → Sample overlap between occasions
  - Inclusion/exclusion of units in the sample on distinct survey rounds
- **Panel**: set of sampling units that join and leave the sample at the same time

BRAZILIAN NATIONAL HOUSEHOLD SURVEY (PNADC)

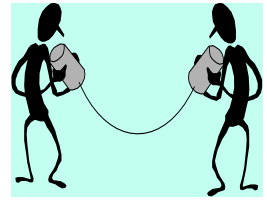
- Publishes official unemployment figures since 2016
- Two-stage cluster design – census enumeration areas are PSUs and households are SSUs
- Rotating panel survey with a partially overlapping sample of households – rotation pattern 1-2(5)
- Planned sample overlap between quarters: 80% of households
- Each household is interviewed once every quarter
- National estimates are released monthly (based on rolling quarterly data), and subnational estimates are published quarterly

CONTEXT

- **Users have been calling for:**
  - estimates based solely on a single-month sample
  - greater frequency of subnational releases
- Single-month estimates are needed to monitor the labour market after the COVID-19 pandemic for state levels
- PNADC sample size big enough for monthly national estimates but not for state-level (**more disaggregation and higher frequency)**
- Alternative data sources as potential data for producing official statistics **(integration of survey and other data sources)**

PNADC TIME SERIES

- Time series of a repeated survey with sample overlap
- Rotation pattern affects the correlation between survey estimates over time
- Observed series are subjected to sampling errors
- The sampling errors are correlated over time due to sample overlap



In the beginning….

Scott and Smith (1974); Binder and Hidiroglou (1988); Binder and Dick (1989) ; Tiller (1989); Pferffermann,

MODELLING TIME SERIES FROM REPEATED SURVEYS

Usual Approach $\hat{y}_t = T_t + S_t + I_t$

Standard time series procedures fail to account for the effect of the sampling error autocorrelation

Signal Extraction $\hat{y}_t = \theta_t + e_t$ Signal + Sampling Error

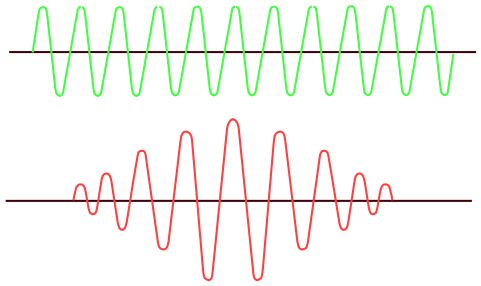
$$\theta_t = T_t + S_t + I_t$$

$\hat{y}_t$ is the design unbiased survey estimate

θ_t is the unknown population quantity

e_t is the sampling error

SIGNAL EXTRACTION APPROACH: **COMBINES TWO MODELS**



- One to describe the evolution of the signal population quantity over time (**signal**) $\{\theta_t\}$
- One to represent the time series relationship between the sampling errors of the survey estimators (**noise**) $\{e_t\}$

MODELLING PROCEDURE

- Formulate time series models for the **signal** θ_t and the **noise** e_t
- Combine the models using a state-space formulation
- The models contemplate the survey design
- Estimate unobserved model components using the Kalman Filter

In the beginning….

Binder and Dick (1990); Pferffermann (1991); Tiller (1992); Pfeffermann, Feder and Signorelli (1998)

TIME SERIES MODEL-BASED ESTIMATORS

Flexibility to meet historical demands and new challenges

- Estimation of trend and seasonality
- Measurement of discontinuities due to survey redesigns
- Production of labour force indicators based solely on the cases surveyed in the reference month (instead of rolling quarters)
- Production of small area estimates
- Estimation of effects due to higher volatility
- Incorporation of auxiliary and alternative data sources, such as big data
- Nowcasting

STATE-SPACE MODEL FOR UNEMPLOYMENT RATE ($\hat{y}_t$)

$\hat{y}_t$: design-based estimate at month t

Signal extraction: $\hat{y}_t = \theta_t + e_t$

Unobserved components of unknown population quantity: $\theta_t = T_t + S_t + I_t \qquad I_t \sim N(0, \sigma_I^2)$

Trend: $T_t = T_{t-1} + R_{t-1}$

$$R_t = R_{t-1} + \eta_{R,t}$$

$$\eta_{R,t} \sim N(0, \sigma_R^2)$$

Seasonal Component

$$S_t = \sum_{l=1}^{\frac{s}{2}=6} S_{l,t} + \eta_{S,t}$$

$$\eta_{S,t} \sim N(0, \sigma_S^2)$$

MODEL FOR PNADC SAMPLING ERROR (e_t)

Sampling error e_t: $e_t = \hat{c}_t \tilde{e}_t$

$\hat{c}_t$: standard error of design-based estimates

PNADC rotation pattern 1-2(5)

Each household is interviewed once every quarter

(model is specified according to sample overlap)

$$\tilde{e}_t = \phi \tilde{e}_{t-3} + \eta_{\tilde{e},t}, \quad \eta_{\tilde{e}} \sim N\left(0, \sigma^2_{\tilde{e}}\right)$$

PROPOSED MODEL FOR PNADC

$$\{\hat{y}_t\} \quad \begin{cases} \hat{y}_t = \theta_t + e_t \end{cases}$$

$$\{\theta_t\} \quad \begin{cases} \theta_t = T_t + S_t + I_t, & I_t \sim N(0, \sigma_I^2) \\ T_t = T_{t-1} + R_{t-1}, & \\ R_t = R_{t-1} + \eta_{R,t}, & \eta_{R,t} \sim N(0, \sigma_R^2) \\ S_t = \sum_{l=1}^{\frac{s}{2}=6} S_{l,t} + \eta_{S,t}, & \eta_{S,t} \sim N(0, \sigma_S^2) \end{cases}$$

$$\{e_t\} \quad \begin{cases} e_t = c_t \tilde{e}_t & \\ \tilde{e}_t = \phi_3 \tilde{e}_{t-3} + \eta_{e,t}, & \eta_e \sim N(0, \sigma_e^2 \cong 1) \end{cases}$$

MODEL BASED ESTIMATES FOR PNADC

Signal Extraction $\hat{y}_t = \theta_t + e_t$

$$\theta_t = T_t + S_t + I_t$$

$\hat{y}_t$ is the design unbiased survey estimate

θ_t is the unknown population quantity

e_t is the survey error

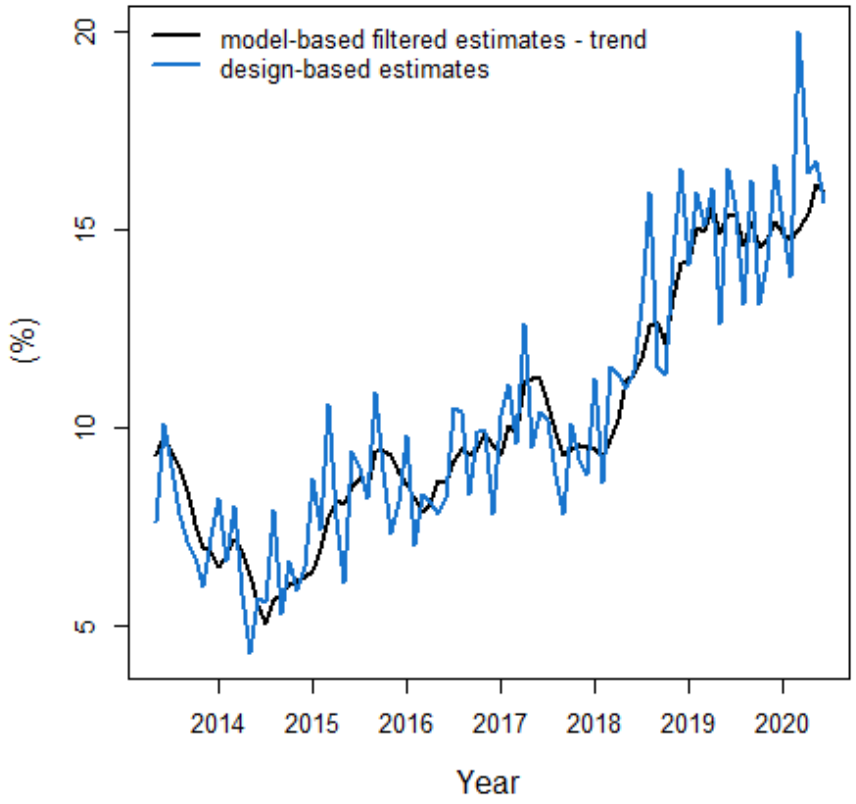
seasonally adjusted series

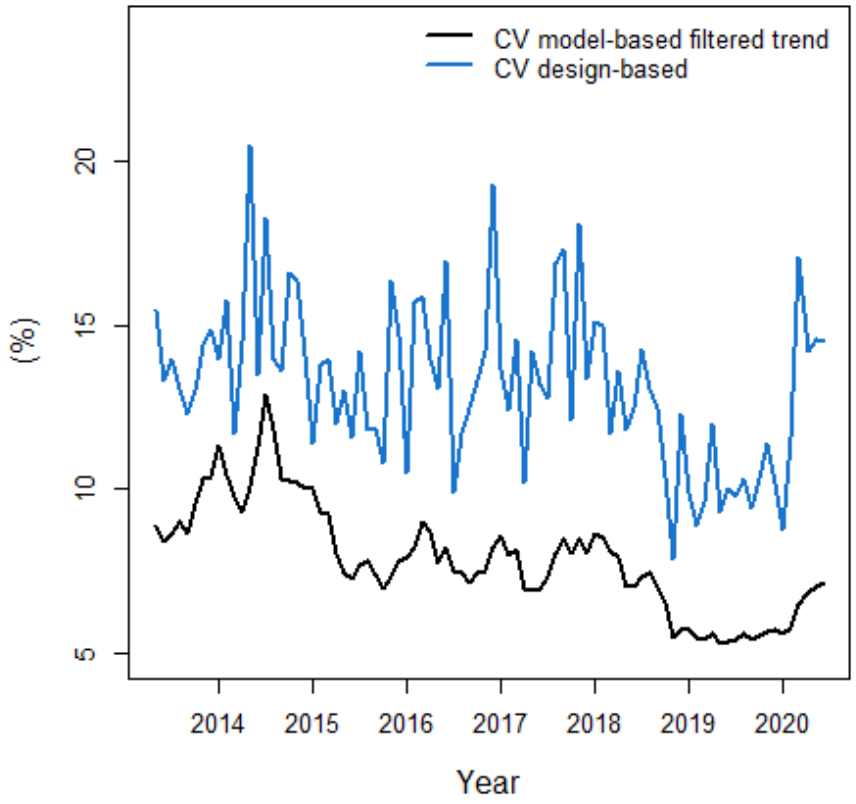
$\theta_t^s = T_t + I_t$

signal: $\theta_t^* = T_t + S_t$

RESULTS

Unemployment rate design-based and model-based (trend) estimates, and coefficients of variation – Roraima





Single month sample size ~ 1,000 households

MODEL-BASED SINGLE MONTH ESTIMATES

ORIGINAL ARTICLE

Single-month unemployment rate estimates for the Brazilian Labour Force Survey using state-space models

Caio Gonçalves, Luna Hidalgo, Denise Silva, Jan van den Brakel

First published: 20 November 2022 | https://doi.org/10.1111/rssa.12914

Users have been calling for estimates based solely on a single-month sample, and for a greater frequency of subnational releases

REGIONAL MULTIVARIATE MODEL

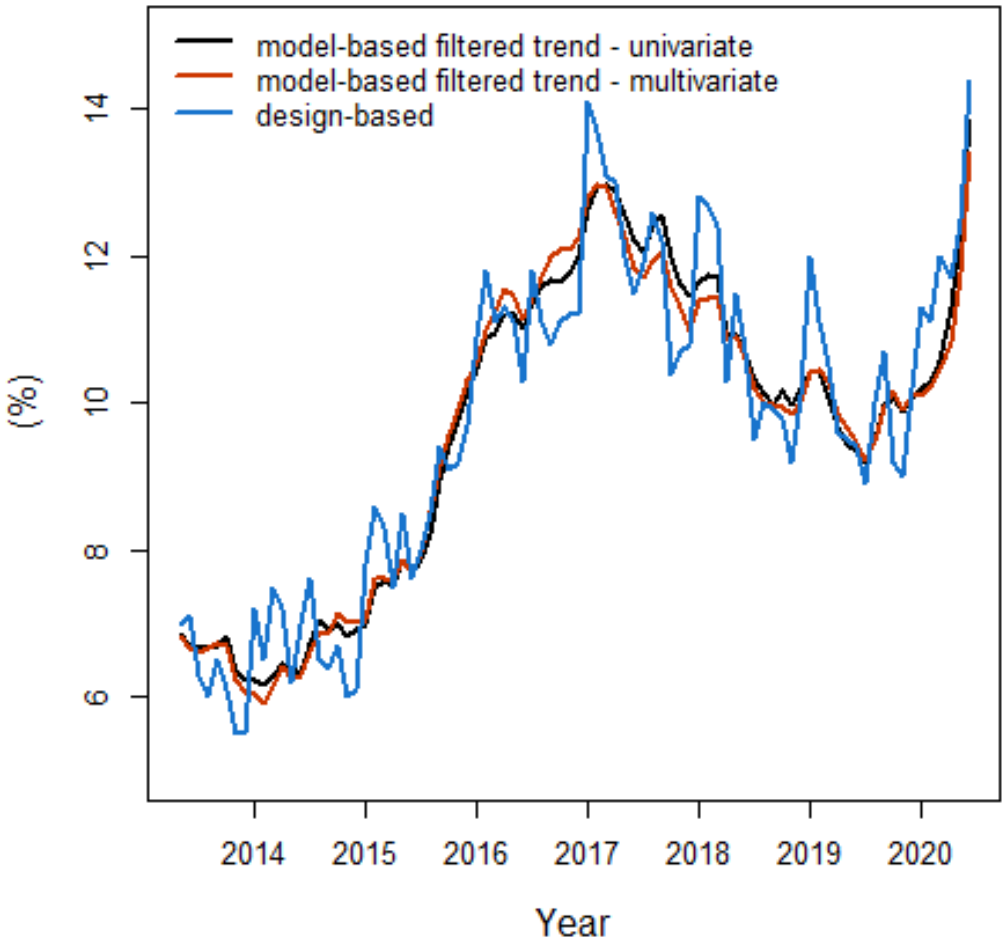
$\hat{y}_{j,t}$: design-based estimate for unemployment rate at month t in the state j

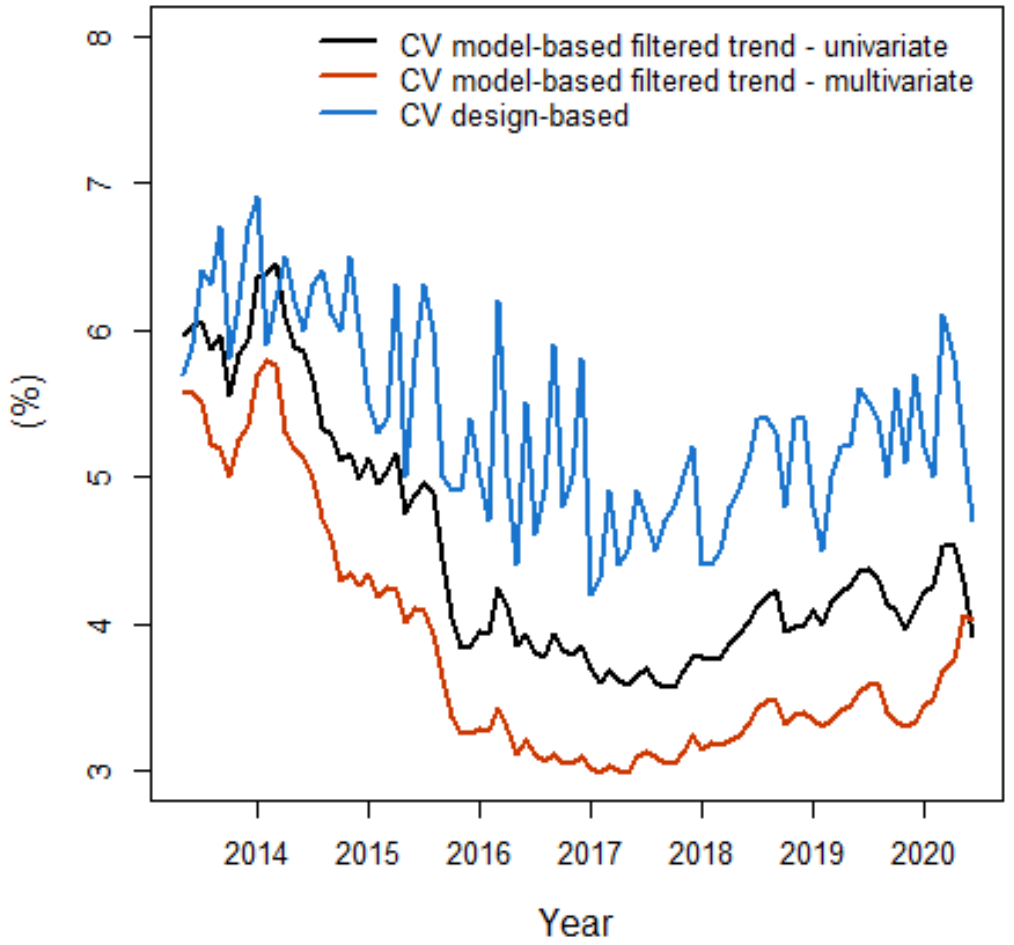
$$\begin{pmatrix} \hat{y}_{1,t} \\ \vdots \\ \hat{y}_{J,t} \end{pmatrix} = \begin{pmatrix} \theta_{1,t} \\ \vdots \\ \theta_{J,t} \end{pmatrix} + \begin{pmatrix} e_{1,t} \\ \vdots \\ e_{J,t} \end{pmatrix} \qquad j = 1, \dots, J$$

Borrowing strength from time and space

Models can be used for Small Area Estimation

DESIGN-BASED, TREND MODEL-BASED (UNIVARIATE AND MULTIVARIATE) ESTIMATES, AND COEFFICIENTS OF VARIATION MINAS GERAIS





REMARKS ON TIME SERIES MODEL-BASED ESTIMATES

- Model-based approach produces trend and seasonally adjusted series taking into account the sampling error
- The advantages of abandoning the use of three-month rolling estimates are wide
- Opportunity to produce monthly estimates for states with acceptable precision
- Regional multivariate models presented advantages

MULTIVARIATE TIME SERIES MODELS TO INTEGRATE SURVEY DATA WITH BIG DATA

Model-based Single-month Unemployment Estimates for the Brazilian Labour Force Survey Incorporating Google Trends Data

- Common trend models to combine survey data and Google Trends time series

$$\begin{pmatrix} \hat{y}_t \\ \boldsymbol{x_t} \end{pmatrix} = \begin{pmatrix} \theta_t \\ \hat{\boldsymbol{\Lambda}}\, \boldsymbol{f_t} \end{pmatrix} + \begin{pmatrix} e_t \\ \boldsymbol{u_t} \end{pmatrix}$$

- $\boldsymbol{x_t}$: Google Trends series
- $\boldsymbol{f_t}$: factors obtained via principal components analysis of $\boldsymbol{x_t}$

Google Trends

- Provides a series of word queries for several countries that can be specific for states or provinces
- The word queries are grouped into categories using a natural language classification engine, such as health, employment, sports, travel, etc.
- Search volume is normalized on a scale from 0-100, where the maximum (100) represents the query's highest point considering a specific beginning and end (GOOGLE, 2022)

Searched words?

Examples....

trabalho	work
vaga	vacancy
vaga de	vacancy of
vaga de emprego	job opportunity
vaga de empregos	job vacancy
vaga de estagio	internship vacancy
vaga de trabalho	job vacancy
vaga emprego	employment vacancy
vagas	vacancies
oportunidades de trabalho	job opportunities
primeiro emprego	first job
processo do trabalho	work process
procuro emprego	I am looking for a job

Use of alternative data through modelling

Multivariate time series models that integrate survey data with big data to produce:

- precise estimates of monthly unemployment figures
- nowcast estimates (since Google Trends series are available one month before the direct single-month unemployment estimates)

Common trend models to combine survey data and Google Trends time series

Gonçalves, C. C. S. Produção de indicadores do mercado de trabalho com modelos de séries temporais de pesquisas repetidas. Tese de doutorado, ENCE, IBGE, 2023.

TIME SERIES MODEL BASED APPROACH

- Flexible and powerful method to produce reliable and precise model-based estimates
- Models can borrow strength from different survey occasions (in time), areas, as well as incorporate administrative or alternative data sources
- Models already tested and/or implemented by BLS-US, Statistics Netherlands, Statistics Canada, ONS-UK, ABS

CHALLENGES

- **Model based approach is tailor-made**
- Multipurpose small area estimation (Chandra and Chambers, 2006)

  https://eprints.soton.ac.uk/38464/
- Consider small area estimation plans when designing a survey

  Sample size calculation for small-area estimation (Longford, 2006)

  https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X20060019259
- Design surveys to leverage data integration
- Intruders also like data integration…

  Statistical Disclosure Control (Templ, 2017) is essential

REFERENCES

Binder, R. and Dick, J. (1989) Modelling and estimation for repeated surveys. *Survey Methodology*, 15, 29–45.

Boonstra, H. J. and van den Brakel, J. A. (2019) Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 395–425.

Pfeffermann, D. and Tiller, R. (2006) Small area estimation with state space models subject to benchmark constraints. *J. Am. Statist. Ass.*, 101, 1387–1397.

Scott, A. J. and Smith, T. M. F. (1974) Analysis of repeated surveys using time series methods. *J. Am. Statist. Ass.*, 69, 674–678.

Scott, A. J., Smith, T. M. F. and Jones, R. G. (1977) The application of time series methods to the analysis of repeated surveys. Int. Statist. Rev., 45, 13–28.

Templ, M. Statistical Disclosure Control for Microdata: Methods and Applications in R, 2017.

Tiller, R.B (1992), “Time Series Modeling of Sample Survey Data From the U.S. Current Population Survey,” Journal of Official Statistics, 8, pp149-166